Real-time scalable object detection

Andrej Karpathy
karpathy@cs.stanford.edu
Joint project with CS229 (with approval from course instructors)

Abstract

Real-time, scalable, multi-view object detection is an active area of research, particularly in robot vision community. An efficient template-based object detection algorithm has recently been proposed [1] that utilizes both color and depth information, and works on texture-less objects. However, the approach scales linearly with the number of objects and views. This project explores a variation of the algorithm that uses a fixed dictionary of templates instead of linearly large, sparse templates, which will allow us to use efficient approximate nearest neighbour methods for efficient matching instead of an exhaustive linear search.

1. Introduction

[1] presents a method for detecting objects that works in real-time, under heavy clutter, does not require a time-consuming training stage, and can handle untextured objects. The algorithm accomplishes this goal by computing feature templates of each object from various views, and storing them in memory. During test time, an efficient matching algorithm compares every patch of an image against the stored templates to detect objects in a scene. The proposed feature templates can leverage both color information and depth information, gathered by sensors such as Kinect.

The problem with [1] is that it slows down linearly with each new object or object view learned. In addition, the features are sparse so they don't fit into existing scaling algorithms such as Locality Sensitive Hashing, or approximate nearest neighbours.

2. Proposed Approach

This project will attempt to address these shortcomings as follows. Instead of maintaining a library of (large, sparse) templates for every new object/view, a set of N random feature templates will first be generated as a dictionary. For every new object/view, instead of storing the feature we will only store its similarity to every one of our N dictionary templates. The resulting N-dimensional vector will be hashed into M-dimensional space ($M < N$) using an efficient feature hashing algorithm such as WTA hash [2], and stored in memory. Since the resulting codes are relatively low-dimensional and dense, existing efficient approximate nearest-neighbour techniques such as Locality Sensitive Hashing [3] can be used for retrieval.

The input data will consist of RGBD views of a set of 5-30 objects obtained using a turntable. The algorithm will learn a model for each object/view using the proposed method above. Evaluation will be carried out on turntable and also on cluttered scenes "in the wild", and will consist of correctness measures (such as True Positive / False Positive rates), analysis of sensitivity to occlusions, and complexity measures (in frames per second, as compared to the previous system). Time permitting, I will explore approaches for online/semi-supervised learning with this framework, as one of the strengths of the approach is that it is very simple to add training data without having to retrain the system.

An existing (messy, incomplete) implementation of the previous method is available in OpenCV/C++. I will have to generate my own data, which will involve camera calibration, data gathering using Kinect, and data preprocessing (such as depth segmentation for objects of interest). I will implement the entire new version of the algorithm (described above) on top of the existing code base. Implementation of Locality Sensitive Hashing is available in FLANN library [4].

References

[1] S. Hinterstoisser, S. Holzer , C. Cagniart, S. Ilic, K. Konolige, N. Navab, V. Lepetit Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes (Oral), IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, November 2011.

[2] "The Power of Comparative Reasoning", Jay Yagnik, Dennis Strelow, David Ross, Ruei-Sung Lin, International Conference on Computer Vision(2011).

[3] "Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions" (by Alexandr Andoni and Piotr Indyk). Communications of the ACM, vol. 51, no. 1, 2008, pp.

117-122.

[4] Marius Muja and David G. Lowe, ”Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration”, in International Conference on Computer Vision Theory and Applications (VISAPP’09), 2009

[5] On random weights and unsupervised feature learning, Andrew Saxe, Pangwei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh and Andrew Y. Ng. In Proceedings of the Twenty-Eighth International Conference on Machine Learning, 2011